Algoritmo de búsqueda de secuencias cis-regulatorias basado en el análisis del incremento de la información mediante la divergencia de Rényi

Joan Maynou, Montserrat Vallverdú, Pere Caminal y Alexandre Perera

Abstract— **La regulación de la expresión génica es un proceso fundamental en el desarrollo y funcionamiento de todo organismo vivo. Es un proceso altamente regulado que involucra la unión de una proteína llamada factor de transcripción (TF) a una secuencia específica (secuencia de unión). En este contexto, la detección de las secuencias regulatorias es fundamental para entender mejor la regulación génica. El principal objetivo de este trabajo es proponer una metodología basada en el incremento de la información para la detección de secuencias regulatorias. Dicha metodología asume que existe una correlación entre las posiciones de los puntos de unión la cuál es caracterizada mediante la divergencia de Rényi. Esta metodología ha sido aplicada en la búsqueda de diferentes factores de transcripción para varios organismos: *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster* y *Homo sapiens*. Los resultados obtenidos han sido visualizados mediante las correspondientes curvas ROC (Receiver Operating Characteristic) y comparados con MEME (Multiple expectation-maximization for motif elecitation) y ITEME (Information Theory for Motif Estimation) mediante un paquete en R llamado MEET (Motif Elements Estimation Toolkit).**

I. INTRODUCCIóN

Un gen es una unidad de ácido desoxirribonucleico (ADN) que contiene la información necesaria para la síntesis, mediante la regulación génica, de un ácido ribonucleico (ARN) o de una secuencia de aminoácidos. Dicha secuencia, también conocida como proteína, es un elemento esencial para el funcionamiento celular. La síntesis de una secuencia de aminoácidos se inicia con la transcripción génica. Proceso en el cuál se crea el ácido ribonucleico (RNA) a partir del ácido desoxirribonucleico (ADN). El primer paso en el proceso de control de la transcripción, es la unión entre una o varias proteínas con una región específica de la secuencia de ADN. Dichas proteínas, conocidas como factores de transcripción (TF), reconocen secuencias específicas de ADN llamadas puntos de unión de los factores de transcripción (TFBS) o secuencias cis-regulatorias. El estudio y la identificación de las secuencias cis-regulatorias es importante para entender y clarificar las redes regulatorias [1]. La identificación de los TFBS es extremadamente difícil debido a las características inherentes de los TFBS. Concretamente, estas secuencias son extremadamente cortas (de 5 a 20 pares de bases) y presentan una gran variabilidad. Debido a estas características, es difícil establecer una secuencia consensus. Como resultado, el número de falsos positivos es alto respecto a la sensitividad. Por lo tanto, cualquier método computacional para la búsqueda de secuencias de unión tiene que considerar las características de los TFBS para poder reducir el número de falsos positivos y aumentar la sensitividad [2].

Recientemente, diferentes aproximaciones computacionales han sido propuestas para la detección de secuencias de unión. Según el método considerado, se clasifican en: determinista, numérico o probabilístico. Los algoritmos basados en un método probabilístico consideran la frecuencia de aparición de $\{A, C, G, T\}$ en cada posición de la secuencia de unión mediante una matriz de pesos (PWM). Dichos modelos asumen la independencia entre las posiciones de la secuencia cis-regulatoria. Diferentes experimentos han demostrado la existencia de cierta dependencia entre las posiciones de los puntos de unión [2]. Algunos autores han propuestos diferentes aproximaciones para incorporar la dependencia entre las posiciones de los motivos en la matriz PWM: modelo mixto [3], cadenas de Markov de orden m^{th} y redes Bayesianas. A partir de dichas aproximaciones, se han desarrollado diferentes algoritmos para mejorar la detección de las secuencias cis-regulatorias como MEME/MAST [4]. MEME (Multiple expectation-maximization for motif elicitation) está basado en el principio de máxima verosimilitud [4].

Por otra parte, la teoría de la información ha sido aplicada en el campo de la genética para la visualización y caracterización de la información contenida en un conjunto de secuencias alineadas mediante las correspondientes secuencias logos y los perfiles de entropía [5]. Además de dichas contribuciones, se han establecido las bases para el uso de las medidas de la teoría de la información para la detección de las secuencias regulatorias. Concretamente, la primera contribución consiste en un detector basado en la entropía de Rényi. Dicha medida, considera la independencia entre las posiciones de la secuencia cis-regulatoria [6]. Esta aproximación está incluida en el paquete de R MEET con el nombre de ITEME (Entropía).

En este artículo, se propone una generalización del método usado para la detección de secuencias cis-regulatorias basada en el incremento de la información de un conjunto de secuen-

Este trabajo se ha realizado con el soporte del Ministerio Español de Educación y Ciencia mediante el programa de la Ramón y Cajal y TEC2010-20886-C02-02 y el CIBER-BBN.

J. Maynou, M. Vallverdú, P. Caminal y A. Perera son del Dep. ESAII, Centre de Recerca en Enginyeria Biomèdica (CREB), Universitat Politècnica de Catalunya (UPC), Barcelona, Pau Gargallo, 5, 08028 Barcelona, España. http://www.creb.upc.es, http://www.upc.edu. e-mail: joan.maynou, montserrat.vallverdu, pere.caminal, alexandre.perera@upc.edu

J. Maynou, M. Vallverdú, P. Caminal y A. Perera como miembros de CIBER de Bioingeniería, Biomateriales y Nanomedicina. http://www.isciii.es/htdocs/redes/ciber.jsp

cias alineadas mediante el uso de la divergencia de Rényi. Dicha aproximación asume la existencia de correlación entre las posiciones de las secuencias cis-regulatorias. Los resultados han sido comparados contra MEME/MAST y ITEME (Entropía) mediante el paquete en R MEET [1].

II. Materiales y Métodos

A. Método

Dado un conjunto de secuencias cis-regulatorias alineadas, la información total contenida en dicho conjunto, puede ser computada mediante una medida de incertidumbre, como por ejemplo la divergencia de Rényi.

A partir de la información inicial del conjunto de entrenamiento, la búsqueda de nuevas secuencias en una secuencia de ADN se basa en el incremento de la información del conjunto inicial cuando se añade una nueva secuencia. Concretamente, si la secuencia que se añade difiere del set de entrenamiento, se produce un incremento en la información total. En cambio, si la nueva secuencia es cercana al conjunto, la información total no variará de forma significativa. Dicho incremento es medido mediante la divergencia de Rényi la cuál considera la dependencia entre las posiciones de la secuencia de unión y permite, mediante el parámetro α, parametrizar el número de puntos de unión correlacionados y su amplitud. El valor óptimo de α corresponde al compromiso entre el ruido que se añade y la pérdida de información en la matriz de divergencia y depende del factor de transcripción.

En este artículo, se propone el uso de la generalización de la divergencia de Kullback-Leibler [7] para medir la correlación entre posiciones de la secuencia cis-regulatoria. Dicha generalización se conoce como divergencia de Rényi o α-Divergencia. La divergencia de Rényi es una divergencia paramétrica la cuál depende del parámetro α, también conocido como el orden de Rényi. La probabilidad conjunta de cada par de símbolos está modulada, enfatizando o reprimiendo dichos valores, según los valores de α [6]. Si el valor de α disminuye, la probabilidad de ocurrencia de cada par de símbolos aumenta. En cambio, si el valor de α aumenta, la probabilidad decrece. Entonces, un detector paramétrico, basado en un modelo de dependencia K_2, puede ser construido a través de esta medida. Además, la sensibilidad del detector está modulada por el valor de α.

Para cada factor de transcripción (ver tabla I), el número de secuencias disponibles en la base de datos es pequeña. Por lo tanto, el detector ha sido caracterizado mediante leave one out cross validation (LOOCV). Cada secuencia individual es utilizada como una secuencia test de un clasificador de entrenamiento con las restantes $n-1$ secuencias, donde n es el número de secuencias. Cada nuevo conjunto de secuencias de entrenamiento es realineado mediante el alineamiento múltiple de secuencias (MSA). Los resultados se han obtenidos para una secuencia genómica de los organismos eucariotas considerados (ver tabla II).

[1] http://sisbio.recerca.upc.edu/R/MEET.5.1.tar.gz.

TABLA I
RESUMEN DE LOS RECONOCEDORES ANALIZADOS

Organismo	Reconocedor	Base	Secuencias Alineadas
Mus musculus	*Mycn*	31	6
Rattus norvegicus	*CREB1*	12	16
Drosophila melanogaster	*VIS*	34	6
Homo sapiens	*ELK1*	28	16

TABLA II
RESUMEN DE LAS SECUENCIAS GENÓMICAS

Reconocedor	Secuencia Genómica	Rango
Mycn	$EP07119(+)MmIgk'MPC11$	$(-1000, 500)$
CREB1	$EP24038(+)RnmyosinLC3_fP2$	$(-1000, 500)$
VIS	$EP17014(+)DmsnRNAU1$	$(-1000, 500)$
ELK1	$EP74078(+)HsRPS9P2+$	$(-1000, 500)$

B. Divergencia de Rényi

La divergencia de Kullback-Leibler (KL) es una medida de incertidumbre que cuantifica, en bits, la proximidad de dos distribuciones de probabilidad P y Q [7]. La divergencia de Rényi es una divergencia paramétrica que puede ser considerada como la generalización de la divergencia de Kullback-Leibler. La divergencia de Rényi de orden α para dos variables discretas, X y Y, con N posibles estados $(X_1, X_2, \cdots, X_i, \cdots, X_N)$ y $(Y_1, Y_2, \cdots, Y_i, \cdots, Y_N)$, está definida como,

$$D_\alpha(X;Y) = \frac{1}{\alpha - 1} log_2 \sum_{i=1}^{N} P_i^\alpha Q_i^{1-\alpha} \tag{1}$$

donde, las variables X y Y son los nucleótidos en dos posiciones diferentes. La divergencia de Rényi es no-negativa para todas las $\alpha > 0$ y converge a la divergencia de Kullback-Leibler cuando α tiende a 1.

$$\lim_{\alpha \to 1} D_\alpha(X;Y) = \sum_N \sum_N P_i log_2 \left(\frac{P_i}{Q_i} \right) \tag{2}$$

C. Descripción de la base de datos

Un conjunto de secuencias alineadas con evidencia de unión es necesario para la búsqueda de nuevas secuencias cis-regulatorias. Estas secuencias provienen de diferentes organismo eucariotas (ver Tabla I). Para cada organismo de estudio se ha considerado un factor transcripción el cuál está caracterizado por su estructura y por su estrategia de interacción con las secuencias cis-regulatorias, Tabla I. La base de datos ha sido obtenida de Jaspar [8], `http://jaspar.genereg.net/`. Los resultados han sido calculados a partir de una secuencia genómica de cada organismo. Dichas secuencias han sido obtenidas de Eukaryotic Promoter Database (EPD) [9], Tabla II.

D. Detección de Motivos

Dada una matriz de secuencias de unión alineadas, la correlación entre las posiciones es calculada mediante la divergencia de Rényi. El grado de correlación entre dichas posiciones es parametrizado mediante el orden Rényi (α).

El estudio ha sido realizado para valores de α comprendidos entre 0 y 2. Se define una función para evaluar el incremento en la información entre la matriz de entrenamiento y la matriz de entrenamiento cuando la secuencia candidata es añadida al conjunto. Está función se define como,

$$\eta = [\gamma * (R * R_t)^{1/2}]^{-1}; \gamma = |D_\alpha - D_\alpha^s| \qquad (3)$$

donde, D_α es la matriz α-divergencia del conjunto de secuencias alineadas, D_α^s es la matriz α-divergencia considerando la matriz de entrenamiento más la secuencia candidata, R es la redundancia del conjunto de secuencias alineadas y R^t es la redundancia traspuesta. Si la secuencia de estudio es una verdadera secuencia de unión, la dependencia entre las posiciones se mantendrá aproximadamente constante, $\gamma \sim 0$. En cambio, cuando la secuencia candidata es una secuencia aleatoria, la dependencia entre posiciones disminuirá y γ aumentará. Por lo tanto, mediante el análisis del incremento en la información, se puede discriminar entre una secuencia aleatoria y una secuencia cis-regulatoria o de unión. El algoritmo desarrollado, basado en el criterio descrito anteriormente, es el siguiente:

1) Estudio preeliminar sobre la correlación entre las posiciones de las secuencias de unión computada mediante la divergencia de Rényi.
2) Corrección del efecto de muestra finita [10], [11].
3) Consideración de las posiciones significativas. En dichas posiciones, se calcula la probabilidad conjunta para cada posible estado de dos símbolos.
4) Calculo de la divergencia de Rényi.
5) Se añade la secuencia candidata y se calcula la divergencia Rényi para el conjunto de matriz de entrenamiento más secuencia de estudio.
6) A partir de la función definida en la ecuación (3), se calcula el incremento en la información cuando se añade la secuencia candidata al conjunto de entrenamiento.

III. Resultados

Para diferentes secuencias cis-regulatorias y para diferentes valores de α se ha obtenido la área bajo la curva ROC (Receiver Operating Characteristic), ver Tabla III. Se ha comparado el detector contra MEME/MAST [4] y ITEME (Entropía) [6]. El detector que produzca una mejor área bajo la superficie convexa (AUC) será el mejor sistema de aprendizaje.

Se puede observar que el detector basado en α-divergencia tiene un mejor comportamiento que los demás detectores, mejorando los resultados obtenidos mediante ITEME (Entropía) y MEME/MAST. Dado un TFBS, el número de verdaderos y falsos positivos depende del parámetro α considerado. El mejor parámetro α es seleccionado para la detección según el criterio de coste establecido entre los verdaderos y falsos positivos y la máxima área bajo la curva ROC.

Dado el valor óptimo de α, la diferencia entre las poblaciones por cada TFBS y el método se visualiza en la tabla III. Se observa que las poblaciones son diferentes según el método y los TFBS utilizados. Básicamente, los grados de dispersión y la simetría en los datos depende del grado de conservación de las posiciones de la secuencia de unión. Para un conjunto de posiciones de la secuencia de unión conservadas, el grado de dispersión en los datos es bajo y la simetría es alta (e.g. *Mus musculus* Myscn). A medida que disminuye la conservación en las posiciones de la secuencia de unión, el grado de dispersión aumenta y la simetría disminuye.

Finalmente, en la Fig. 1 se visualiza la salida para los métodos basados en la entropía y en la divergencia para diferentes factores de transcripción con su correspondiente valor óptimo de α. En el espacio de salida, se observa que los valores están divididos en dos espacios: en verdaderas y falsas secuencias de unión. Secuencias con alta conservación y alta correlación, tienen valores altos de divergencia y entropía. Estas secuencias son potenciales secuencias de unión. En cambio, secuencias con baja conservación y baja correlación, tienen valores pequeños de divergencía y entropía. Dichas secuencias no corresponden a secuencias de unión. A medida que la conservación y la correlación decrece, el número de verdaderas secuencias de unión decrece y crece el número de falsas secuencias de unión. Por lo tanto, la medida óptima de detección de secuencias de unión corresponde a la combinación de ambas medidas, entropía y divergencia.

IV. Conclusión

Se ha presentado una metodología para la detección de secuencias cis-regulatorias basada en en la medida del incremento de la información en un conjunto de secuencias alineadas cuando se añade una nueva secuencia. Se ha utilizado una medida no-lineal paramétrica basada en la teoría de la información llamada divergencia de Rényi o α-divergencia. Dicha medida calcula la distancia entre las posiciones de las secuencias de unión. La parametrización permite modular el número de posiciones de la secuencia de unión correlacionadas y su amplitud. La α-divergencia aporta información adicional entre la correlación de las posiciones de la secuencia de unión con respecto a la información mutua. El parámetro α óptimo depende de las características de las secuencias de unión del factor de transcripción. Este parámetro tiene que ser ajustado para cada conjunto de secuencias mediante la validación cruzada. Este algoritmo ha sido aplicado en detección para los factores de transcripción *Mycn*, *CREB1*, *VIS* y *ELK1*. Los resultados obtenidos mejoran la detección de secuencias cis-regulatorias basada en la entropía de Rényi y MEME/MAST. Además, se ha observado que la combinación de ambas medidas, entropía y divergencía, es la medida óptima de detección de secuencias de unión. Para futuros estudios se integrará en un sólo detector la conservación de los puntos de unión de los factores de transcripción (medidas de incertidumbre paramétricas) y la dependencia entre posiciones de unión (medidas de divergencia paramétrica)

TABLA III

ÁREA BAJO LA CURVA ROC

	Entropía			Divergencia			MEME/MAST	
TFBS	α	AUC	Error	α	AUC	Error	AUC	Error
Mycn	0.5	0.99817	0.00862	0.5	**0.99933**	0.00345	0.99872	0.00905
CREB1	1.5	0.99971	0.00084	0.5	**0.99987**	0.00036	0.99952	0.00142
VIS	0.5	0.93448	0.0849	1.5	**0.99532**	0.01168	0.97874	0.04769
EL1	1.5	0.99341	0.01941	2.0	**0.99398**	0.00358	0.98849	0.02085

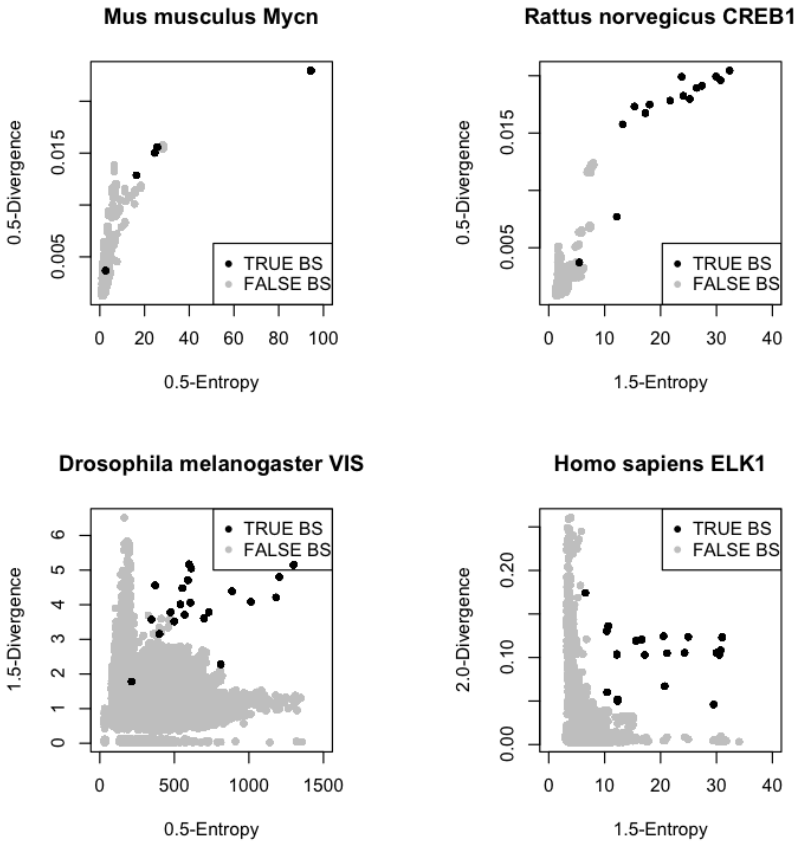

Fig. 1. Izquierda a derecha: Representación gráfica de la salida del detector basado en la entropía de Rényi respecto al detector basado en la divergencia de Rényi para los factores de transcripción Mycn, CREB1, VIS y ELK1.

V. ACKNOWLEDGMENTS

CIBER de Bioingeniería, Biomateriales y Nanomedicina es una iniciativa de ISCIII.

REFERENCES

[1] W. Wei and X.-D. Yu, "Comparative analysis of regulatory motif discovery tools for transcription factor binding sites," *Geno. rot. Bioinfo*, vol. 5, no. 2, pp. 131–142, 2007.

[2] A. Tomovic and E. Oakeley, "Position dependencies in transcription factor binding sites," *Bioinformatics*, vol. 23, no. 8, pp. 933–941, 2007.

[3] Y.Barash, G. Elidean, N. Friedman, and T. Kaplan, "Modeling dependencies in protein-dna binding sites," in *RECOMB*, 2003.

[4] T. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, August 1994, pp. 28–36.

[5] T. D. Schneider and R. M. Stephens, "Sequence logos: A new way to display consensus sequences," *J Mol Biol*, vol. 18, pp. 6097–6100, 1990.

[6] J. Maynou, M. Vallverdu, J. Gallardo-Chacon, P. Caminal, and A. Perera, "Computational detection of transcription factor binding sites using a parametric entropy measure," *IEEE Trans. Information Theroy*, vol. 56, no. 2, pp. 734–741, 2010.

[7] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79–86, 1951.

[8] D. Vlieghe, A. Sandelin, P. D. Bleser, K. Vleminckx, W. Wasserman, F. V. Roy, and B. Lenhard, "A new generacion of jaspar, the open-access repository for transcription factor binding site profiles," *Nucleic Acids Research*, vol. 34 (Database issue), pp. D95–D97, 2006.

[9] C. Schmid, R. Perier, and P. Bucher, "Edp in its twentieth year: towards complete promoter coverage of selected model organisms," *Nucleic Acids Research*, vol. 34, pp. D82–85, 2006.

[10] B. Goebel, Z. Dawy, J. Hagenauer, and J. Mueller, "An approximation to the distribution of finite sample size mutual infomation estimates," pp. 1102–1106, 2005.

[11] P. Kumar, "Generalized relative j-divergence measure and properties," *Int. J. Contemp. Math. Sci*, vol. 13, pp. 597–609, 2006.